Pharmacoepidemiology and Drug Safety

WILEY

COMMENTARY OPEN ACCESS

Metadata for Data dIscoverability aNd Study rEplicability in obseRVAtional Studies (MINERVA): Lessons Learnt From the MINERVA Project in Europe

Rosa Gini[1] | Romin Pajouheshnia[2,3] | Lia Gutierrez[3] | Morris A. Swertz[4] | Eleanor Hyde[4] | Miriam Sturkenboom[5] | Alejandro Arana[3] | Carla Franzoni[3] | Vera Ehrenstein[6] | Giuseppe Roberto[1] | Miguel Gil[7] | Miguel Angel Maciá[7] | Wiebke Schäfer[8] | Ulrike Haug[8,9] | Nicolas H. Thurin[10] | Régis Lassalle[10] | Cécile Droz-Perroteau[10] | Silvia Zaccagnino[11] | Maria Paula Busto[11] | Bas Middelkoop[11] | Karin Gembert[12] | Francisco Sanchez-Saez[13] | Clara Rodriguez-Bernal[13] | Gabriel Sanfélix-Gimeno[13] | Isabel Hurtado[13] | Manuel Barreiro-de Acosta[14] | Beatriz Poblador-Plou[15] | Jonás Carmona-Pírez[15,16] | Antonio Gimeno-Miguel[15] | Alexandra Prados-Torres[15] | Anna Schultze[17] | Ella Jansen[18] | Ron Herings[18] | Josine Kuiper[18] | Igor Locatelli[19] | Janja Jazbar[19] | Špela Žerovnik[19] | Mitja Kos[19] | Steven Smit[20] | Sirje Lind[20] | Andres Metspalu[20] | Stefania Simou[21] | Karin Hedenmalm[21] | Ana Cochino[21] | Paolo Alcini[21] | Xavier Kurz[21] | Susana Perez-Gutthann[3]

[1]Agenzia Regionale di Sanità Della Toscana, Florence, Italy | [2]Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht University, Utrecht, The Netherlands | [3]Department of Epidemiology, RTI Health Solutions, Barcelona, Spain | [4]Department of Genetics, University Medical Centre Groningen, Groningen, The Netherlands | [5]Department of Datascience and Biostatistics, University Medical Centre Utrecht, Utrecht, The Netherlands | [6]Department of Clinical Epidemiology, Aarhus University and Aarhus University Hospital, Aarhus, Denmark | [7]Base de Datos Para la Investigación Farmacoepidemiológica en Atención Primaria, Agencia Española de Medicamentos y Productos Sanitarios, Madrid, Spain | [8]Department of Clinical Epidemiology, Leibniz Institute for Prevention Research and Epidemiology, Bremen, Germany | [9]Faculty of Human and Health Science, University of Bremen, Bremen, Germany | [10]Bordeaux PharmacoEpi, INSERM CIC-P 1401, Université de Bordeaux, Bordeaux, France | [11]European Society for Blood & Marrow Transplantation, Leiden, The Netherlands | [12]Department of Epidemiology, Karolinska Institutet, Stockholm, Sweden | [13]Health Services Research & Pharmacoepidemiology Unit, Foundation for the Promotion of Health and Biomedical Research of the Valencia Region, Valencia, Spain | [14]Spanish Working Group on Crohn's Disease and Ulcerative Colitis, Bilbao, Spain | [15]EpiChron Research Group, Aragon Health Sciences Institute (IACS), IIS Aragón, Miguel Servet University Hospital, Zaragoza, Spain | [16]Technical Advisory Subdirectorate of Information Management (STAGI), Andalusian Health Service (SAS), Seville, Spain | [17]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK | [18]PHARMO Institute for Drug Outcomes Research, Utrecht, The Netherlands | [19]University of Ljubljana, Faculty of Pharmacy, Ljubljana, Slovenia | [20]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia | [21]European Medicines Agency, Amsterdam, The Netherlands

Correspondence: Rosa Gini (rosa.gini@ars.toscana.it)

Received: 22 December 2023 | **Revised:** 30 May 2024 | **Accepted:** 17 July 2024

Funding: This project was funded by the European Medicines Agency (EMA) through the framework contract No. EMA/2017/09/PE/16.

Keywords: catalogue | FAIR | metadata | observational studies | real-world data sources | reproducibility

Research from this project was presented at the 2022 International Conference of Pharmacoepidemiology (ICPE), August 2022. Copenhagen, Denmark: (1) MINERVA: Metadata for Data dIscoverability aNd Study rEplicability in obseRVAtional studies; Pajouheshnia R. et al. Poster No. 163, Publication No. 1297, 28 August 2022; (2) MINERVA: Study Scripts Supporting Multiple Common Data Models; Gini R et al. Poster no. 137, Publication 1182, 28 August 2022; and (3) Data source heterogeneity in multi-database pharmacoepidemiologic studies: an ISPE-sponsored scoping review. DIVERSE Symposium, 26 August 2022.

The views expressed in this article are the personal views of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the regulatory agency/agencies or organizations with which the authors is/are employed/affiliated.

1 | Introduction

In November 2020, the MINERVA (Metadata for data dIscoverability aNd study rEplicability in obseRVAtional studies) project (EUPAS39322) was initiated in response to the Heads of Medicines Agencies–European Medicines Agency (HMA–EMA) joint Big Data Task Force recommendation on "the identification of metadata" for regulatory decision-making on the choice of data source [1–3].

The project primarily aimed at defining a set of metadata and developing a good practice guide describing the metadata and recommendations on the use and sustainability of metadata collection. The second objective was to pilot a proof-of-concept metadata catalogue. The metadata list and the proof-of-concept metadata catalogue were designed to assist investigators, research partners, and other evidence consumers in understanding the data flow in data sources. This understanding supports the whole research process: data source discoverability, study feasibility assessment, study design and execution, and interpretation of study results.

The project's methods and results are fully reported in the accompanying article in this issue of *Pharmacoepidemiology and Drug Safety* [4]. This commentary goes beyond the project or the European setting and focuses on possible general application of the results, lessons learnt, and recommendations.

2 | Applications

2.1 | Use Cases

The metadata list generated by the project was designed to support multiple-use cases, including

- Supporting data access partners (DAPs) in incorporating their knowledge about the data source.
- Supporting discoverability of data source(s) suitable for conducting of a given study.
- Assisting investigators in designing single-database or multi-database studies, and interpreting the results.
- Supporting analysts with programming analyses in a study.
- Helping readers of study reports to interpret results and understand limitations of reproducibility across different studies.
- Allowing institutions that initiate or maintain data source catalogues adopting the same standards to map their metadata list to the MINERVA metadata list (or a subset thereof) and reuse metadata.

Examples of such use cases are reported in the accompanying article by Pajouheshnia et al. [4], in a 2022 EMA draft Guidance document [5], and in the final report of the MINERVA study [6]. To best support such use cases, metadata would need to cover both data sources and registrations of studies. Therefore, the metadata list was designed to address both domains.

2.2 | Implementations of the Metadata List

The metadata list is a standalone tool, available in a ready-to-use spreadsheet form (Pajouheshnia et al. [4], supplementary material 2), which provides a standard for describing real-world data sources and studies. As an example, submissions to the ongoing special section of this journal, *Real-World Data Sources for Pharmacoepidemiologic Research*, could complete and submit (a subset of) the metadata list.

The open-source software Molgenis was used during the project to implement the metadata list in a catalogue prototype that could be accessed and updated online. The same software can be used by others to build additional implementations and collect their metadata [7]. However, in principle, other software can be used to implement the metadata list as an online catalogue.

Several implementations of this metadata list, or subsets thereof, in catalogues of real-world data sources are described in the accompanying the article by Pajouheshnia et al. [4].

3 | Populating and Maintaining Catalogues: Lessons Learnt and Recommendations

This section collects lessons learnt, structured under several subsections leading to the final considerations on sustainability.

3.1 | A Community of FAIR Metadata Catalogues

Several past initiatives (e.g., GRiP, ADVANCE/AIRR, EMIF [8], IMI-ConcePTION, Vaccine Monitoring Collaboration for Europe-VAC4EU [9]) coexist in Europe that involve a structural collection of metadata on data sources. In February 2024, the EMA launched the HMA–EMA Catalogues of real-world data sources and studies, which enhance both the European Union electronic Register of Post-Authorisation Studies (EU PAS Register) [10] and the ENCePP (European Network of Centres for Pharmacoepidemiology and Pharmacovigilance) Resources Database [1]. The challenge identified in the HMA–EMA Joint Big Data Taskforce Phase II report has been to create sustainable and FAIR data sources and metadata catalogues [11]. The acronym FAIR refers to four principles: findable, accessible, interoperable, and reusable [1]. Systematic interoperability among catalogues might reduce metadata entry frequency, increase quality, foster perceived value, and increase sustainability.

Findable and *accessible* means that the catalogues should be accessible to the public as an online tool. In the MINERVA project, we observed a lack of clarity over whether the experts involved in creating the metadata (DAPs) were authorized to release the metadata content in public, and clarification is needed. Indeed, normally other organizations ("data originators") generate the data for other purposes, while the pharmacoepidemiology expertise of DAPs is needed to identify strengths and limitations for secondary use. We advocate for DAPs to be able to make their scientific assessments publicly available in future catalogues.

Based on the experience of the MINERVA proof-of-concept catalogue, we identified three key requirements for *interoperability* across catalogues.

First, persistent identifiers (PIDs) for the key objects, and in particular institutions, data sources and data banks need to be created and maintained by an authority [12]. While there are multiple authorities whose mandate is compatible with supporting PIDs for institutions, there is a need for authorities specifically interested in data reuse to take responsibility to support PIDs for data sources and data banks. The MINERVA recommendations highlighted that EMA was in a position of providing such a service, and the new HMA–EMA Catalogues of real-world data sources do indeed support PIDs for data sources. This may possibly prove to be an additional added value in the context of the European Health Data Space ecosystem [13]. It must be noted that, while the focus of the HMA–EMA Catalogues is Europe, they also accept registration of non-European data sources [14].

Second, metadata lists need to be mapped one to another: the metadata list of the MINERVA project is a candidate master data model for this purpose.

Third, global ontologies need to be established for many metadata: it was a consistent finding by the MINERVA project that for many metadata global ontologies are lacking or inconsistent. This finding is consistent with the result of a recent scoping review that provides a foundation for guidance on reporting data source diversity by the International Society of Pharmacoepidemiology [15]. The scientific community is actively working on this, and alignment should be sought.

Those three elements also support *reusability* of metadata already collected, thus avoiding duplication of the effort required to maintain high-quality and up-to-date metadata, ultimately supporting sustainability. A recent review issued a similar recommendation and extended it to the case of data primarily generated for research [16].

3.2 | Catalogue Population and Maintenance: Qualitative Metadata

Based on our experiences in the MINERVA proof-of-concept catalogue, we recommend that metadata for new catalogue entries is collected in an interview, where the data expertise of the DAP is met with the expertise on the metadata list of a metadata expert. In line with this, we recommend that a Catalogue Quality Office (CQO) is maintained by the funders of a catalogue, that should also be supported by automated checks (for instance of the format of the entries). Update and maintenance of the catalogue requires both engagement from the DAPs and effort from the CQO (see recommendations below under the *sustainability* header). Version control is essential, and content of each version should be attributed to an author. A robust authentication system—for example, ELIXIR Authentication and Authorisation Infrastructure (AAI) (https://elixir-europe.org/services/compute/aai)—should be adopted.

If metadata describing *data sources* and *studies* are designed to be interoperable within (or between) catalogues, versions of the metadata describing the data source at the moment of the study could be created. This would allow knowledge about a data source gained during a study to be collected and help investigators to use the catalogue throughout a study.

Finally, a constant alignment with global alliances and initiatives for standards in labeling/annotating metadata should be sought, to keep the metadata list up to date and interoperable. In particular, alignment with PIDs and global ontologies should be sought.

3.3 | Catalogue Population and Maintenance: Quantitative Metadata

In the final MINERVA metadata list, only simple quantitative metadata are included, such as yearly population size per gender and age. During the project, the case of computation of more complex quantitative metadata was also assessed. They are metadata whose computation requires epidemiological expertise, such as the occurrence of a condition in a data source population. First, computation of such metadata requires that entrance and exit from the data source's underlying population are defined. The MINERVA list includes metadata describing the criteria for entrance and exit from the data source's underlying population. If dates are not collected primarily in the data source, surrogate dates must be defined and their limitations must be recorded to support interpretation. Second, the condition must be defined, considering the information available in the data source, which also may come with limitations that must be recorded (e.g., drug proxies may be used alongside diagnostic codes to detect some conditions in some data sources). Third, the distribution of important covariates in the underlying population must be considered to support comparability. The recommendation is that quantitative metadata that require epidemiological expertise are recorded in study protocols and signed off by a DAP for each data source.

Finally, a distributed approach to the computation of quantitative metadata is recommended for efficiency and transparency [17]. Multiple common data models can be supported by the central procedure, as demonstrated during the MINERVA project.

A separate recommendation included in the final MINERVA good practice guide is also pertinent to this topic: that the ability of cross-linking each data source to the studies that have been conducted in the past, including reports of results, may be of help to understand suitability of the data source for a new study. This recommendation suggests that, in some sense, results from older studies may play themselves the role of "complex quantitative metadata" for the data source. This is indeed possible in the HMA–EMA Catalogues of real-world data sources and studies, which provide a direct link between a study and the data sources used in the study.

3.4 | Legal Context

Metadata catalogues must adhere to data protection regulations. For Europe, the General Data Protection Regulation poses important requirements, including defining and communicating the purpose and use of the catalogue as well as the duration of the availability of metadata; obtaining approval for metadata to

be reused, reedited, and published under an appropriate license prior to entry; ensuring a technical option to delete metadata; and establishing appropriate measures to protect the privacy of institutions and individuals.

In Europe, the legislation around secondary use of data for research purposes is evolving, especially around the European Health Data Space [13], an initiative of the European Commission.

3.5 | Sustainability

Set-up and maintenance of a catalogue is costly, requiring effort and engagement by the CQO and by the DAPs. In the final study report of the MINERVA study, section 3.9, an exercise of quantification of effort was conducted, both for the implementation and for the maintenance phase [6]. For example, during maintenance, a yearly update of overall metadata content of a data source was estimated to require a person-day for each data source, to be provided by the corresponding DAP.

This could be partly mitigated if the task of entering metadata in a publicly available FAIR catalogue is included among the tasks of funded studies. Future catalogues should be designed with sustainability at the forefront. This would allow investigators to include catalogue maintenance among the activities funded by studies. This strategy is included in the draft document "Reflection paper on use of real-world data to generate world evidence in non-interventional studies" that the EMA recently issued for public consultation. In the draft, a recommendation is made to marketing authorization holders (MAHs), applicants, and concerned stakeholders to register in the HMA–EMA Catalogues of data sources the data sources used in non-interventional studies submitted to regulators (most of these studies should also be registered in the catalogue of studies); if the data source is already registered, it would be appropriate to update the information if the last update was performed more than 12 months ago. This provision may be included in the contractual agreement between the MAH or applicant and the DAP, as relevant.

As mentioned above, interoperability among FAIR catalogues would then support sustainability.

4 | Conclusion

We listed a series of recommendations to address the challenges of creating and maintaining a FAIR metadata catalogue, also addressing the critical challenge of sustainability. Collaboration across research networks and stakeholders within Europe and across the world is needed. To achieve sustainability, interoperability in a community of FAIR metadata catalogues should be enabled.

Author Contributions

Authors worked collaboratively and contributed to the work performed throughout the duration of the project and were also involved in the preparation or review of the manuscript.

Ethics Statement

The research conducted did not involve human subjects and as such ethics reviews or approvals were not required.

Consent

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The information that supports the findings of this project, including main deliverables and the metadata list, is available through the European Union Post-Authorisation Studies Register (EU PAS Register) at https://www.encepp.eu/encepp/viewResource.htm?id=49345.

References

1. M. Wilkinson, M. Dumontier, I. Aalbersberg, et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* 15, no. 3 (2016): 160018, https://doi.org/10.1038/sdata.2016.18.

2. HMA–EMA, "Priority Recommendations of the HMA-EMA Joint big Data Task Force," HMA-EMA Big Data Steering Committee Group, accessed May 14, 2024, https://www.ema.europa.eu/en/documents/other/priority-recommendations-hma-ema-joint-big-data-task-force_en.pdf.

3. HMA–EMA, "Big Data Steering Group. Workplan," accessed May 14, 2024, https://www.hma.eu/fileadmin/dateien/HMA_joint/00-_About_HMA/03-Working_Groups/Big_Data/2020_09_HMA-EMA_Big_Data_SG_Workplan.pdf.

4. R. Pajouheshnia, R. Gini, L. Gutierrez, et al., "Metadata for Data dIscoverability aNd Study rEplicability in obseRVAtional Studies: Definition and Recommendations of Use From the MINERVA Project in Europe," *Pharmacoepidemiology and Drug Safety* (2024), https://doi.org/10.1002/pds.5871.

5. European Medicines Agency, "Good Practice Guide for the Use of the Metadata Catalogue of Real-World Data Sources V 1.0," accessed December 19, 2023, https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/good-practice-guide-use-metadata-catalogue-real-world-data-sources_en.pdf.

6. EU PAS Register, "European Network of Centres for Pharmacoepidemiology and Pharmacovigilance. MINERVA Deliverable 9, Final Good Practice Guide for Metadata Collection for Real-World Data Sources," accessed December 19, 2023, https://www.encepp.eu/encepp/openAttachment/studyResult/45315;jsessionid=WSmD0hsat-YOzKOXWFJbyVHKPvgt8kiUQrE3kkq3TasxFiQrYe-w!-1399792416.

7. M. A. Swertz, M. Dijkstra, T. Adamusiak, et al., "The MOLGENIS Toolkit: Rapid Prototyping of Biosoftware at the Push of a Button," *BMC Bioinformatics* 11, no. 12 (2010): S12, https://doi.org/10.1186/1471-2105-11-S12-S12.

8. J. L. Oliveira, A. Trifan, and L. A. Bastião Silva, "EMIF Catalogue: A Collaborative Platform for Sharing and Reusing Biomedical Data," *International Journal of Medical Informatics* 126 (2019): 35–45.

9. VAC4EU, "Toolbox Catalogue," accessed December 19, 2023, https://vac4eu.org/catalogue/.

10. X. Kurz, S. Perez-Gutthann, and ENCePP Steering Group, "Strengthening Standards, Transparency, and Collaboration to Support Medicine Evaluation: Ten Years of the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP)," *Pharmacoepidemiology and Drug Safety* 27, no. 3 (2018): 245–252.

11. Heads of Medicines Agencies, European Medicines Agency, "HMA-EMA Joint Big Data Taskforce Phase II Report: Evolving Data-Driven Regulation," accessed October 11, 2022, https://www.ema.europa.eu/en/documents/other/hma-ema-joint-big-data-taskforce-phase-ii-report-evolving-data-driven-regulation_en.pdf.

12. European Commission, "Directorate-General for Research and Innovation," in *A Persistent Identifier (PID) Policy for the European Open Science Cloud (EOSC)*, eds. M. Hellström, A. Heughebaert, R. Kotarski, et al. (Luxembourg: Publications Office of the EU, 2020), https://data.europa.eu/doi/10.2777/926037.

13. EHDS, "The European Health Data Space," accessed December 19, 2023, https://www.european-health-data-space.com/.

14. European Medicines Agency, "RWD Catalogues: Support. FAQ No. 57," accessed May 14, 2024, https://catalogues.ema.europa.eu/support.

15. R. Gini, R. Pajouheshnia, H. Gardarsdottir, et al., "Describing Diversity of Real World Data Sources in Pharmacoepidemiologic Studies: The DIVERSE Scoping Review," *Pharmacoepidemiology and Drug Safety* 33, no. 5 (2024): e5787.

16. M. Swertz, E. van Enckevort, J. L. Oliveira, et al., "Towards an Interoperable Ecosystem of Research Cohort and Real-World Data Catalogues Enabling Multi-Center Studies," *Yearbook of Medical Informatics* 31, no. 1 (2022): 262–272.

17. R. Gini, M. C. J. Sturkenboom, J. Sultana, et al., "Different Strategies to Execute Multi-Database Studies for Medicines Surveillance in Real-World Setting: A Reflection on the European Model," *Clinical Pharmacology and Therapeutics* 108, no. 2 (2020): 228–235.